SD-TSIA 211 – Optimization for Machine Learning

Pascal Bianchi, Olivier Fercoq, Anne Sabourin

November 15, 2017

Contents

- $\min_{x \in \mathbb{R}^d} f(x)$ ← convexe.
- Analyse convexe.
- Algo de gradient & variants.
- Dualité.

Chapter 1

Introduction: Optimization, machine learning and convex analysis

1.1 Optimization problems in Machine Learning

Most of Machine Learning algorithms consist in solving a minimization problem. In other words, the output of the algorithm is the solution (or an approximated one) of a minimization problem. In general, non-convex problems are difficult, whereas convex ones are easier to solve. Here, we are going to focus on convex problems.
First, let's give a few examples of well-known issues you will have to deal with in supervised learning :

Example 1.1.1 (Least squares, simple linear regression or penalized linear regression).

(a) Ordinary Least Squares:

$$\min_{x \in \mathbb{R}^p} \|Zx - Y\|^2, Z \in \mathbb{R}^{n \times p}, Y \in \mathbb{R}^n$$

(b) Ridge :

$$\min_{x \in \mathbb{R}^p} \|Zx - Y\|^2 + \lambda \|x\|_2^2,$$

(c) Lasso :

$$\min_{x \in \mathbb{R}^p} \|Zx - Y\|^2 + \lambda \|x\|_1,$$

Example 1.1.2 (Linear classification).
The data consists of a training sample $\mathcal{D} = \{(z_1, y_1), \ldots, (z_n, y_n)\}$, $y_i \in \{-1, 1\}$, $z_i \in \mathbb{R}^p$, where the z_i's are the data's *features* (also called *regressors*), whereas the y_i's are the labels which represent the class of each observation i. The sample is obtained by independent realizations of a vector $(Z, Y) \sim P$, of unknown distribution P. Linear classifiers are linear functions defined on the *feature space*, of the kind:

$$h : z \mapsto \text{sign}(\langle x, z \rangle + x_0) \qquad (x \in \mathbb{R}^p, x_0 \in \mathbb{R})$$

A classifier h is thus determined by a vector $\mathbf{x} = (x, x_0)$ in $\mathbb{R}^{p+1}$. The vector x is the normal vector to an hyperplane which separates the space into two regions, inside which the predicted labels are respectively "$+1$" and "-1".

The goal is to learn a classifier which, in average, is not wrong by much: that means that we want $\mathbb{P}(h(Z) = Y)$ to be as big as possible.
To quantify the classifier's error/accuracy, the reference loss function is the '0-1 loss':

$$L_{01}(\mathbf{x}, z, y) = \begin{cases} 0 & \text{if } -y\,(\langle x, z\rangle + x_0) \leq 0 \quad (h(z) \text{ and } y \text{ of same sign}), \\ 1 & \text{otherwise .} \end{cases}$$

In general, the implicit goal of machine learning methods for supervised classification is to solve (at least approximately) the following problem:

$$\min_{\mathbf{x}\in\mathbb{R}^{p+1}} \frac{1}{n}\sum_{i=1}^{n} L_{0,1}(\mathbf{x}, z_i, y_i) \tag{1.1.1}$$

i.e. to minimize the *empirical risk*.
As the cost L is not convex in $\mathbf{x}$, the problem (1.1.1) is *hard*. Classical Machine learning methods consist in minimizing a function that is similar to the objective (1.1.1): the idea is to replace the cost 0-1 by a *convex substitute*, and then to add a penalty term which penalizes "complexity" of x, so that the problem becomes numerically feasible. More precisely, the problem to be solved numerically is

$$\min_{x\in\mathbb{R}^p, x_0\in\mathbb{R}} \sum_{i=1}^{n} \varphi(-y_i(\,x^\top z_i + x_0)) + \lambda\,\mathcal{P}(x), \tag{1.1.2}$$

where $\mathcal{P}$ is the penalty and φ is a convex substitute to the cost 0-1.
Different choices of penalties and convex subsitutes are available, yielding a range of methods for supervised classification :

- For $\varphi(u) = \max(0, 1+u)$ (Hinge loss), $\mathcal{P}(x) = \|x\|^2$, this is the SVM:

$$\min_{x\in\mathbb{R}^p, x_0\in\mathbb{R}} \sum_{i=1}^{n} \max(0, 1 - y_i(\,x^\top z_i + x_0)) + \lambda\|x\|^2.$$

- In the separable case (*i.e.* when there is a hyperplane that separates the two classes), introduce the "convex indicator function" (also called *characteristic function*),

$$\iota_A(x) = \begin{cases} 0 & \text{if } x \in A, \\ +\infty & \text{if } x \in A^c, \end{cases} \qquad (A \subset \mathcal{X})$$

  and set

$$\varphi(u) = \iota_{\mathbb{R}^-}(u).$$

  The solution to the problem is the maximum margin hyperplane:

$$\begin{aligned} &\min_{x\in\mathbb{R}^p, x_0\in\mathbb{R}} \sum_{i=1}^{n} \iota(-y_i(\,x^\top z_i + x_0)) + \lambda\|x\|^2 \\ &= \min_{x\in\mathbb{R}^p, x_0\in\mathbb{R}} \{\lambda\|x\|^2, \text{ st: } -y_i(\,x^\top z_i + x_0) \leq 0, \forall i, 1 \leq i \leq n\} \end{aligned}$$

- For $\varphi(u) = \log(1+\exp(u))$ (Logistic loss) and $\mathcal{P}(x) = \|x\|^2$, this is the logistic regression:

$$\min_{x\in\mathbb{R}^p, x_0\in\mathbb{R}} \sum_{i=1}^{n} \log(1 + \exp(-y_i(\,x^\top z_i + x_0))) + \lambda\|x\|^2.$$

To summarize, the common denominator of all these versions of example 1.1.2 is as follows:

- The risk of a classifier x is defined by $J(x) = \mathbb{E}(L(x, D))$. We are looking for x which minimizes J.
- $\mathbb{P}$ is unknown, and so is J. However, $D \sim \mathbb{P}$ is available. Therefore, the approximate problem is to find:
$$\hat{x} \in \arg\min_{x \in \mathcal{X}} J_n(x) \hat{=} \frac{1}{n} \sum_{i=1}^{n} L(x, d_i)$$
- The cost L is replaced by a convex surrogate L_φ, so that the function $J_{n,\varphi} = \frac{1}{n} \sum_{i=1}^{n} L_\varphi(x, d_i)$ convex in x.
- In the end, the problem to be solved, when a convex penalty term is incorporated, is
$$\min_{x \in \mathcal{X}} J_{n,\varphi}(x) + \lambda \mathcal{P}(x). \tag{1.1.3}$$

In the remaining of the course, the focus is on that last point: how to solve the convex minimization problem (1.1.3)?

1.2 General formulation of the problem

In this course, we only consider optimization problems which are defined on a finite dimension space $\mathcal{X} = \mathbb{R}^n$. These problems can be written, without loss of generality, as follows:

$$\begin{gathered} \min_{x \in \mathcal{X}} f(x) \\ s.t. \text{ (such that / under constraint that)} \\ g_i(x) \leq 0 \text{ for } 1 \leq i \leq p, \quad F_i(x) = 0 \text{ for } 1 \leq i \leq m\,. \end{gathered} \tag{1.2.1}$$

The function f is the *target function* (or *target*),
the vector
$$C(x) = (g_1(x), \ldots, g_p(x), F_1(x), \ldots, F_m(x))$$
is the (functional) constraint vector.
The region
$$K = \{x \in \mathcal{X} : \; g_i(x) \leq 0, \; 1 \leq i \leq p, \quad F_i(x) = 0, \; 1 \leq i \leq m\,\}$$
is the set of *feasible* points.

- If $K = \mathbb{R}^n$, this is an *unconstrained* optimization problem.
- Problems where $p \geq 1$ and $m = 0$, are referred to as *inequality contrained* optimization problems.
- If $p = 0$ and $m \geq 1$, we speak of *equality contrained* optimization.
- When f and the contraints are regular (differentiable), the problem is called *differentiable* or *smooth*.
- If f or the contraints are not regular, the problem is called *non-differentiable* or *non-smooth*.

- If f and the contraints are convex, we have a *convex* optimization problem (more details later).

Solving the general problem (1.2.1) consists in finding

- a minimizer $x^* \in \arg\min_K f$ (if it exists, *i.e.* if $\arg\min_K f \neq \emptyset$),
- the *value* $f(x^*) = \min_{x \in K} f(x)$,

We can rewrite the constrained problem as an unconstrained problem, thanks to the infinite indicator function ι introduced earlier. Let's name g and (resp) F the vectors of the inequality and (resp) equality contraints.
For $x, y \in \mathbb{R}^n$, we write $x \preceq y$ if $(x_1 \leq y_1, \ldots, x_n \leq y_n)$ and $x \not\preceq y$ otherwise. The problem (1.2.1) is equivalent to :

$$\min_{x \in E} f(x) + \iota_{g \preceq 0, F=0}(x) \tag{1.2.2}$$

Let's notice that, even if the initial problem is smooth, the new problem isn't anymore !

1.3 Algorithms

Approximated solutions Most of the time, Problem (1.2.1) cannot be analytically solved. However, numerical algorithms can provide an approximate solution. Finding an ϵ-approximate solution (**ϵ-solution**) consists in finding $\hat{x} \in K$ such that, if the "true" minimum x^* exists, we have

- $\|\hat{x} - x^*\| \leq \epsilon$,
  and/or
- $|f(\hat{x}) - f(x^*)| \leq \epsilon$.

"Black box" model A standard framework for optimization is the **black box**. That is, we want to optimize a function in a situation where:

- The target f is not entirely accessible (otherwise the problem would already be solved !)
- The algorithm does not have any access to f (and to the constraints), except by successive calls to an *oracle* $\mathcal{O}(x)$.
  Typically, $\mathcal{O}(x) = f(x)$ (0-order oracle) or $\mathcal{O}(x) = (f(x), \nabla f(x))$ (1-order oracle), or $\mathcal{O}(x)$ can evaluate higher derivative of f ($\geq$ 2-order oracle).
- At iteration k, the algorithm only has the information $\mathcal{O}(x_1), \ldots, \mathcal{O}(x_k)$ as a basis to compute the next point x_{k+1}.
- The algorithm stops at time k if a criterion $T_\epsilon(x_k)$ is satisfied: the latter ensures that x_k is an ϵ-solution.

Performance of an algorithm Performance is measured in terms of computing resources needed to obtain an approximate solution.
This obviously depends on the considered problem. A **class of problems** is:

- A class of target functions (regularity conditions, convexity or other)
- A condition on the starting point x_0 (for example, $\|x - x_0\| \leq R$)
- An oracle.

Definition 1.3.1 (oracle complexity)**.** The **oracle complexity** of an algorithm $\mathcal{A}$, for a class of problems C and a given precision ϵ, is the minimal number $N_{\mathcal{A}}(\epsilon)$ such that, for all objective functions and any initial point $(f, x_0) \in C$, we have:

$$N_{\mathcal{A}}(f, \epsilon) \leq N_{\mathcal{A}}(\epsilon)$$

where : $N_{\mathcal{A}}(f, \epsilon)$ is the number of calls to the oracle that are needed for $\mathcal{A}$ to give an ϵ-solution. The oracle complexity, as defined here, is a *worst-case* complexity. The computation time depends on the oracle complexity, but also on the number of required arithmetical operations at each call to the oracle. The total number of arithmetic operations to achieve an ϵ-solution in the worst case, is called *arithmetic complexity*. In practice, it is the arithmetic complexity which determines the computation time, but it is easier to prove bounds on the oracle complexity .

1.4 Preview of the rest of the course

A natural idea to solve general problem (1.2.1) is to start from an arbitrary point x_0 and to propose the next point x_1 in a region where f "has a good chance" to be smaller.
If f is differentiable, one widely used method is to follow "the line of greatest slope", i.e. move in the direction given by $-\nabla f$.
What's more, if there is a local minimum x^*, we then have $\nabla f(x^*) = 0$. So a similar idea to the previous one is to set the gradient equal to zero.
Here we have made implicit assumptions of regularity, but in practice some problems can arise.

- Under which assumptions is the necessary condition '$\nabla f(x) = 0$' sufficient for x to be a local minimum?
- Under which assumptions is a local minimum a global one?
- What if f is not differentiable?
- How should we proceed when E is a high-dimensional space?
- What if the new point x_1 leaves the admissible region K?

The appropriate framework to answer the first two questions is convex analysis. The lack of differentiability can be bypassed by introducing the concept of *subdifferential*. *Duality* methods solve a problem related to ((1.2.1)), called *dual problem*. The dual problem can often be easier to solve (*ex:* if it belongs to a space of smaller dimension). Typically, once the dual solution is known, the primal problem can be written as a unconstrained problem that is easier to solve than the initial one. For example, *proximal* methods can be used to solve constrained problems.

To go further ...
A panorama in Boyd and Vandenberghe (2009), chapter 4, more rigor in Nesterov (2004)'s introduction chapter (easy to read !).

Chapter 2

Elements of convex analysis

Throughout this course, the functions of interest are defined on a subset of $\mathcal{X} = \mathbb{R}^n$. We will also need a Euclidean space $\mathbf{E}$, endowed with a scalar product denoted by $\langle \cdot, \cdot \rangle$ and an associated norm $\| \cdot \|$. In practice, the typical setting is $\mathbf{E} = \mathcal{X} \times \mathbb{R} = \mathbb{R}^{n+1}$.
Notations: For convenience, the same notation is used for the scalar product in $\mathcal{X}$ and in $\mathbf{E}$. If $a \leq b \in \mathbb{R} \cup \{-\infty, +\infty\}$, $(a, b]$ is an interval open at a, closed at b, with similar meanings for $[a, b)$, (a, b) and $[a, b]$.

N.B The proposed exercises include basic properties for you to demonstrate. You are strongly encouraged to do so ! The exercises marked with * are less essential.

2.1 Convexity

Definition 2.1.1 (Convex set)**.** A set $K \subset \mathbf{E}$ is **convex** if

$$\forall (x, y) \in K^2, \forall t \in [0, 1], \quad t\,x + (1 - t)\,y \in K.$$

Exercise 2.1.1.

1. Show that a ball, a vector subspace or an affine subspace of $\mathbb{R}^n$ are convex.
2. Show that any intersection of convex sets is convex.

In constrained optimization problems, it is useful to define cost functions with value $+\infty$ outside the admissible region. For all $f : \mathcal{X} \to [-\infty, +\infty]$, the *domain* of f, denoted by $\mathrm{dom}(f)$, is the set of points x such that $f(x) < +\infty$.
A function f is called **proper** if $\mathrm{dom}(f) \neq \emptyset$ (i.e $f \not\equiv +\infty$) and if f *never* takes the value $-\infty$.

Definition 2.1.2. Let $f : \mathcal{X} \to [-\infty, +\infty]$. The **epigraph of** f, denoted by $\mathrm{epi}\, f$, is the subset of $\mathcal{X} \times \mathbb{R}$ defined by:

$$\mathrm{epi}\, f = \{(x, t) \in \mathcal{X} \times \mathbb{R} : \quad t \geq f(x)\ \}.$$

Beware: the "ordinates" of points in the epigraph always lie in $(-\infty, \infty)$, by definition.

Definition 2.1.3 (Convex function)**.** $f : \mathcal{X} \to [-\infty, +\infty]$ is **convex** if its epigraph is convex.

Proposition 2.1.1. *A function $f : \mathcal{X} \to [-\infty, +\infty]$ is convex if and only if*

$$\forall (x, y) \in \mathcal{X}^2,\ \forall t \in (0, 1), \quad f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y).$$

Proof. Assume that f satisfies the inequality. Let (x,u) and (y,v) be two points of the epigraph: $u \geq f(x)$ and $v \geq f(y)$. In particular, $(x,y) \in \mathrm{dom}(f)^2$. Let $t \in]0,1[$. The inequality implies that $f(tx+(1-t)y) \leq tu+(1-t)v$. Thus, $t(x,u)+(1-t)(y,v) \in \mathrm{epi}(f)$, which proves that $\mathrm{epi}(f)$ is convex.
Conversely, assume that $\mathrm{epi}(f)$ is convex. If $x \notin \mathrm{dom}\, f$ or $y \notin \mathrm{dom}\, f$, the inequality is trivial. So let us consider $(x,y) \in \mathrm{dom}(f)^2$. For (x,u) and (y,v) two points in $\mathrm{epi}(f)$, and $t \in [0,1]$, the point $t(x,u)+(1-t)(y,v)$ belongs to $\mathrm{epi}(f)$. So, $f(t(x+(1-t)y) \leq tu+(1-t)v$.

- If $f(x)$ et $f(y)$ are $> -\infty$, we can choose $u = f(x)$ and $v = f(y)$, which demonstrates the inequality.
- If $f(x) = -\infty$, we can choose u arbitrary close to $-\infty$.Letting u go to $-\infty$, we obtain $f(t(x+(1-t)y) = -\infty$, which demonstrates here again the inequality we wanted to prove. □

Exercise 2.1.2. Show that:

1. If f is convex, then $\mathrm{dom}(f)$ is convex.
2. If f_1, f_2 are convex and $a, b \in \mathbb{R}_+$, then $af_1 + bf_2$ is convex.
3. If f is convex and $x, y \in \mathrm{dom}\, f$, then for all $t \geq 1$, $z_t = x + t(y-x)$ satisfies the inequality $f(z_t) \geq f(x) + t(f(y) - f(x))$.
4. If f is convex, proper, with $\mathrm{dom}\, f = \mathcal{X}$, and if f is bounded, then f is constant.

Exercise 2.1.3. *
Let f be a convex function and x, y in $\mathrm{dom}\, f$, $t \in (0,1)$ and $z = tx + (1-t)y$. Assume that the three points $(x, f(x)), (z, f(z))$ and $(y, f(y))$ are aligned. Show that for all $u \in (0,1)$, $f(u\,x + (1-u)\,y) = u\,f(x) + (1-u)\,f(y)$.

In the following, the **upper hull** of a family $(f_i)_{i\in I}$ of convex functions will play a key role. By definition, the upper hull of the family is the function $x \mapsto \sup_i f_i(x)$.

Proposition 2.1.2. *Let $(f_i)_{i\in I}$ be a family of convex functions $\mathcal{X} \to [-\infty, +\infty]$, with I any set of indices. Then* ***the upper hull of the family*** $(f_i)_{i\in I}$ ***is convex.***

Proof. Let $f = \sup_{i\in I} f_i$ be the upper hull of the family.

(a) $epi f = \bigcap_{i\in I} \mathrm{epi}\, f_i$. Indeed,

$$(x,t) \in \mathrm{epi}\, f \Leftrightarrow \forall i \in I, t \geq f_i(x) \Leftrightarrow \forall i \in I, (x,t) \in \mathrm{epi}\, f_i \Leftrightarrow (x,t) \in \cap_i \mathrm{epi}\, f_i.$$

(b) Any intersection of convex sets $K = \cap_{i\in I} K_i$ is convex (exercice 2.1.1)

(a) and (b) show that $\mathrm{epi}\, f$ is convex, *i.e.* that f is convex. □

Proposition* 2.1.3. *Let $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a* **jointly** *convex function. Then the function*

$$\begin{aligned} f : \mathcal{X} &\to \mathbb{R} \\ x &\mapsto \inf_{y\in\mathcal{Y}} F(x,y) \end{aligned}$$

is convex.

Proof. Let $u, v \in \mathcal{X}$ and $\alpha \in (0,1)$. We need to show that $f(\alpha u+(1-\alpha)v) \leq \alpha f(u)+(1-\alpha)f(v)$.

$$\begin{aligned}
\alpha f(u) + (1-\alpha) f(v) &= \alpha \inf_{y_u \in \mathcal{Y}} F(u, y_u) + (1-\alpha) \inf_{y_v \in \mathcal{Y}} F(v, y_v) && \text{(definition of } f\text{)} \\
&= \inf_{y_u \in \mathcal{Y}, y_v \in \mathcal{Y}} \alpha F(u, y_u) + (1-\alpha) F(v, y_v) && \text{(separable problems)} \\
&\geq \inf_{y_u \in \mathcal{Y}, y_v \in \mathcal{Y}} F(\alpha u + (1-\alpha)v, \alpha y_u + (1-\alpha) y_v) && \text{(joint convexity of } F\text{)} \\
&= \inf_{y \in \mathcal{Y}} F(\alpha u + (1-\alpha), y) && \text{(change of variable)} \\
&= f(\alpha u + (1-\alpha)v)
\end{aligned}$$

A valid change of variable is $(y, y') = (\alpha y_u + (1-\alpha) y_v, y_v)$. It is indeed invertible since we have $\alpha \in (0,1)$. □

Definition 2.1.4 (Strong convexity)**.** A function f is μ-strongly convex if $f - \frac{\mu}{2}\|\cdot\|^2$ is convex.

Proposition 2.1.4. *A function $f : \mathcal{X} \to [-\infty, +\infty]$ is μ-strongly convex if and only if*

$$\forall (x,y) \in \mathcal{X}^2,\ \forall t \in (0,1), \quad f(tx + (1-t)y) \leq t f(x) + (1-t) f(y) - \frac{\mu}{2} t(1-t)\|x-y\|^2.$$

2.2 Lower semi-continuity

In this course, we will consider functions with infinite values. Such function cannot be continuous. However, some kind of continuity would be very desirable. For infinite-valued convex function, lower semi-continuity is the good generalization of continuity.

Definition 2.2.1 (Reminder: lim inf : **limit inferior**)**.**
The **limit inferior** of a sequence $(u_n)_{n \in \mathbb{N}}$, where $u_n \in [-\infty, \infty]$, is

$$\liminf(u_n) = \sup_{n \geq 0} \Big(\inf_{k \geq n} u_k \Big).$$

Since the sequence $V_n = \inf_{k \geq n} u_k$ is non decreasing, an equivalent definition is

$$\liminf(u_n) = \lim_{n \to \infty} \Big(\inf_{k \geq n} u_k \Big).$$

Definition 2.2.2 (Lower semicontinuous function)**.** A function $f : \mathcal{X} \to [-\infty, \infty]$ is called **lower semicontinuous (l.s.c.) at** $x \in \mathcal{X}$ if for all sequence (x_n) which converges to x,

$$\liminf f(x_n) \geq f(x).$$

The function f is said to be **lower semicontinuous**, if it is l.s.c. at x, for all $x \in \mathcal{X}$.

The interest of l.s.c. functions becomes clear in the next result

Proposition 2.2.1 (epigraphical characterization)**.** *Let $f : \mathcal{X} \to [-\infty, +\infty]$, any function f is l.s.c. if and only if its epigraph is closed.*

Proof. If f is l.s.c., and if $(x_n, t_n) \in \operatorname{epi} f \to (\bar{x}, \bar{t})$, then, $\forall n, t_n \geq f(x_n)$. Consequently,

$$\bar{t} = \liminf t_n \geq \liminf f(x_n) \geq f(\bar{x}).$$

Thus, $(\bar{x}, \bar{t}) \in \operatorname{epi} f$, and $\operatorname{epi} f$ is closed.
Conversely, if f is *not* l.s.c., there exists an $x \in \mathcal{X}$, and a sequence $(x_n) \to x$, such that $f(x) > \liminf f(x_n)$, *i.e.*, there is an $\epsilon > 0$ such that $\forall n \geq 0$, $\inf_{k \geq n} f(x_k) \leq f(x) - \epsilon$. Thus, for all n, $\exists k_n \geq k_{n-1}$, $f(x_{k_n}) \leq f(x) - \epsilon$. We have built a sequence $(w_n) = (x_{k_n}, f(x) - \epsilon)$, each term of which belongs to $\operatorname{epi} f$, and which converges to a limit $\bar{w} = (f(x) - \epsilon)$ which is outside the epigraph. Consequently, $\operatorname{epi} f$ is not closed. □

Lower semi-continuity is a very desirable property for a function we want to optimize thanks to the following proposition.

Proposition 2.2.2. *Let f be a l.s.c function such that $\lim_{\|x\|\to+\infty} f(x) = +\infty$. Then there exists x^* such that $f(x^*) = \inf_{x\in\mathcal{X}} f(x)$.*

Proof. Let $(x_n)_{n\geq 0}$ be a minimizing sequence, that is a sequence of $\mathcal{X}$ such that we have $\lim_{n\to\infty} f(x_n) = \inf_{x\in\mathcal{X}} f(x)$.
Suppose that (x_n) were unbounded. Then there would exist a subsequence $(x_{\phi(n)})$ such that $\lim_{n\to\infty}\|x_{\phi(n)}\| \to +\infty$. By the assumptions on f, this implies that $\lim_{n\to\infty} f(x_n) = +\infty$ which contradicts the fact that $(x_n)_{n\geq 0}$ is a minimizing sequence.
Thus (x_n) is bounded and we can extract from is a subsequence $(x_{\phi(n)})$ converging to, say, x_*. As f is l.s.c., we get $\inf_{x\in\mathcal{X}} f(x) = \lim_{n\to\infty} f(x_{\phi(n)}) = \liminf f(x_{\phi(n)}) \geq f(x^*) \geq \inf_{x\in\mathcal{X}} f(x)$. □

A nice property of the family of l.s.c. functions is its stability with respect to point-wise suprema.

Proposition 2.2.3. *Let $(f_i)_{i\in I}$ a family of l.s.c. functions. Then, the upper hull $f = \sup_{i\in I} f_i$ is l.s.c.*

Proof. Let C_i denote the epigraph of f_i and $C = \operatorname{epi} f$. As already shown (proof of proposition 2.1.2), $C = \cap_{i\in I} C_i$. Each C_i is closed, and any intersection of closed sets is closed, so C is closed and f is l.s.c. □

Exercise 2.2.1. Show that a function f is l.s.c. if and only if its level sets :

$$L_{\leq\alpha} = \{x \in \mathcal{X} : f(x) \leq \alpha\}$$

are closed.
(see, *e.g.*, Rockafellar et al. (1998), Theorem 1.6.)

2.3 Subdifferential

A classical property of convex function is that they are above their tangents. In a multi-dimensional setting, tangents become tangent hyperplanes

Proposition 2.3.1. *Let $f : \mathcal{X} \to (-\infty, \infty]$ be a convex function, differentiable in x. Then for all $y \in \mathcal{X}$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$$

Proof.

$$\begin{aligned}\langle \nabla f(x), y - x\rangle &= \lim_{t\to 0} \frac{f(x + t(y - x)) - f(x)}{t} \\ &= \lim_{t\to 0^+, t\leq 1} \frac{1}{t}(f(ty + (1 - t)x) - f(x))\end{aligned}$$

For $0 \leq t \leq 1$, $f(ty + (1 - t)x) - f(x) \leq tf(y) + (1 - t)f(x) - f(x)$ so
$\langle \nabla f(x), y - x\rangle \leq \lim_{t\to 0^+, t\leq 1} \frac{1}{t}(tf(y) - tf(x)) = f(y) - f(x)$ □

When the function is not differentiable, we generalize the notion of gradient as follows.

Definition 2.3.1 (Subdifferential)**.** Let $f : \mathcal{X} \to [-\infty, +\infty]$ and $x \in \text{dom}(f)$. A vector $\phi \in \mathcal{X}$ is called a **subgradient** of f at x if:

$$\forall y \in \mathcal{X}, \quad f(y) - f(x) \geq \langle \phi, y - x \rangle \ .$$

The **subdifferential** of f in x, denoted by $\partial f(x)$, is the whole set of the subgradients of f at x. By convention, $\partial f(x) = \emptyset$ if $x \notin \text{dom}(f)$.

Interest: Gradient methods in optimization can still be used in the non-differentiable case, choosing a subgradient in the subdifferential.

Proposition 2.3.2. *Let $f : \mathcal{X} \to (-\infty, \infty]$ be a convex function, differentiable in x. Then $\partial f(x) = \{\nabla f(x)\}$*

Proof. If f is differentiable at x, Proposition 2.3.1 shows that $\partial f(x) \neq \emptyset$. Let $\phi \in \partial f(x)$ and $t \neq 0$. Then for all $y \in \text{dom}(f)$, $f(y) - f(x) \geq \langle \phi, y - x \rangle$. Applying this inequality to $y = x + t(\phi - \nabla f(x))$ leads to :

$$\frac{f\left(x + t(\phi - \nabla f(x))\right) - f(x)}{t} \geq \langle \phi, \phi - \nabla f(x) \rangle \ .$$

The left term converges to $\langle \nabla f(x), \phi - \nabla f(x) \rangle$ by definition of the directional derivative. Finally,

$$\langle \nabla f(x) - \phi, \phi - \nabla f(x) \rangle \geq 0,$$

i.e. $\phi = \nabla f(x)$. □

In order to clarify in what cases the subdifferential is non-empty, we need two more definitions:

Definition 2.3.2. A set $A \subset \mathcal{X}$ is called an **affine space** if, for all $(x, y) \in A^2$ and for all $t \in \mathbb{R}$, $x + t(y - x) \in A$. The **affine hull** $\mathcal{A}(C)$ of a set $C \subset \mathcal{X}$ is **the smallest affine space** that contains C.

Definition 2.3.3. Let $C \subset \mathbf{E}$. The **topology relative to** C is a topology on $\mathcal{A}(C)$. The open sets in this topology are the sets of the kind $\{V \cap \mathcal{A}(C)\}$, where V is open in $\mathbf{E}$.

Definition 2.3.4. Let $C \subset \mathcal{X}$. The **relative interior** of C, denoted by $\text{relint}(C)$, is the interior of C for the topology relative to C. In other words, it consists of the points x that admit a neighborhood V, open in $\mathbf{E}$, such that $V \cap \mathcal{A}(C) \subset C$.

Clearly, $\text{int}(C) \subset \text{relint}(C)$. What's more, one can show that if C is convex, then $\text{relint}(C) \neq \emptyset$.

Proposition 2.3.3. *Let $f : \mathcal{X} \to [-\infty, +\infty]$ be a convex function and $x \in \text{relint}(\text{dom}\, f)$. Then $\partial f(x)$ is non-empty.*

Proof. The proof is a bit technical and uses the concept of separating hyperplane Bauschke and Combettes (2011). □

Remark 2.3.1 (the question of $-\infty$ values)**.**
If $f : \mathcal{X} \to [-\infty, +\infty]$ is convex and if $\text{relint}\, \text{dom}\, f$ contains a point x such that $f(x) > -\infty$, then f *never* takes the value $-\infty$. So f is proper.

Exercise 2.3.1. Show this point, using proposition 2.3.3.

Example 2.3.1. The absolute-value function $x \mapsto |x|$ defined on $\mathbb{R} \to \mathbb{R}$ admits as a subdifferential the sign application, defined by :

$$\text{sign}(x) = \begin{cases} \{1\} & \text{if } x > 0 \\ [-1,1] & \text{if } x = 0 \\ \{-1\} & \text{ifi } x < 0\,. \end{cases}$$

Exercise 2.3.2. Determine the subdifferentials of the following functions, at the considered points :

1. In $\mathcal{X} = \mathbb{R}$, $f(x) = \iota_{[0,1]}$, at $x = 0, x = 1$ and $0 < x < 1$.
2. In $\mathcal{X} = \mathbb{R}^2$, $f(x) = \iota_{B(0,1)}$ (closed Euclidian ball), at $\|x\| < 1$, $\|x\| = 1$.
3. In $\mathcal{X} = \mathbb{R}^2$, $f(x_1, x_2) = \iota_{x_1 < 0}$, at x such that $x_1 = 0$, $x_1 < 0$.
4. $\mathcal{X} = \mathbb{R}$,
$$f(x) = \begin{cases} +\infty & \text{if } x < 0 \\ -\sqrt{x} & \text{if } x \geq 0 \end{cases}$$
at $x = 0$, and $x > 0$.
5. $\mathcal{X} = \mathbb{R}^n$, $f(x) = \|x\|$, determine $\partial f(x)$, for any $x \in \mathbb{R}^n$.
6. $\mathcal{X} = \mathbb{R}$, $f(x) = x^3$. Show that $\partial f(x) = \emptyset$, $\forall x \in \mathbb{R}$. Explain this result.
7. $\mathcal{X} = \mathbb{R}^n$, $C = \{y : \|y\| \leq 1\}$, $f(x) = \iota_C(x)$. Give the subdifferential of f at x such that $\|x\| < 1$ and at x such that $\|x\| = 1$.

   Hint: For $\|x\| = 1$:

   - Show that $\partial f(x) = \{\phi : \ \forall y \in C, \langle \phi, y - x \rangle \leq 0\}$.
   - Show that $x \in \partial f(x)$ using Cauchy-Schwarz inequality. Deduce that the cone $\mathbb{R}_+ x = \{t\,x \ : \ t \geq 0\} \subset \partial f(x)$.
   - To show the converse inclusion : Fix $\phi \in \partial f$ and pick $u \in \{x\}_\perp$ (*i.e.*, u s.t. $\langle u, x \rangle = 0$). Consider the sequence $y_n = \|x + t_n u\|^{-1}(x + t_n u)$, for some sequence $(t_n)_n, t_n > 0, t_n \to 0$. What is the limit of y_n?

     Consider now $u_n = t_n^{-1}(y_n - x)$. What is the limit of u_n? Conclude about the sign of $\langle \phi, u \rangle$.

     Do the same with $-u$, conclude about $\langle \phi, u \rangle$. Conclude.
8. Let $f : \mathbb{R}^n \to \mathbb{R}$, differentiable. Show that: f is convex, if and only if
$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0.$$

2.4 Operations on subdifferentials

Until now, we have seen examples of subdifferential computations on basic functions, but we haven't mentioned how to derive the subdifferentials of more complex functions, such as sums or linear transforms of basic ones. A basic fact from differential calculus is that, when all the terms are differentiable, $\nabla(f + g) = \nabla f + \nabla g$. Also, if M is a linear operator, then we have the equality $\nabla(g \circ M)(x) = M^* \nabla g(Mx)$. Under qualification assumptions, these properties are still valid in the convex case, up to replacing the gradient by the subdifferential and point-wise operations by set operations. But first, we need to define operations on sets.

Definition 2.4.1 (addition and transformations of sets)**.** Let $A, B \subset \mathcal{X}$. The Minkowski sum and difference of A and B are the sets

$$A + B = \{x \in \mathcal{X} : \exists a \in A, \exists b \in B,\ x = a + b\}$$
$$A - B = \{x \in \mathcal{X} : \exists a \in A, \exists b \in B,\ x = a + b\}$$

Let $\mathcal{Y}$ another space and M any mapping from $\mathcal{X}$ to $\mathcal{Y}$. Then MA is the image of A by M,

$$MA = \{y \in \mathcal{Y} : \exists a \in A,\ y = Ma\}.$$

Proposition 2.4.1. *Let $f : \mathcal{X} \to (-\infty, +\infty]$, $g : \mathcal{Y} \to (-\infty, \infty]$ two convex functions and let $M : \mathcal{X} \to \mathcal{Y}$ a linear operator.*

$$\forall x \in \mathcal{X}, \partial f(x) + M^* \partial g(Mx) \subseteq \partial(f + g \circ M)(x)$$

Moreover, if $0 \in \operatorname{relint}(\operatorname{dom} g - M \operatorname{dom} f)$, then

$$\forall x \in \mathcal{X}, \partial(f + g \circ M)(x) = \partial f(x) + M^* \partial g(Mx)$$

Proof. We first show that $\partial f(\,\cdot\,) + M^* \partial g(M\,\cdot\,) \subseteq \partial(f + g \circ M)(\,\cdot\,)$. Let $x \in \mathcal{X}$ and let $\phi \in \partial f(x) + M^* \partial g \circ M(x)$, which means that $\phi = u + M^* v$ where $u \in \partial f(x)$ and $v \in \partial g \circ M(x)$. In particular, none of the latter subdifferentials is empty, which implies that $x \in \operatorname{dom} f$ and $x \in \operatorname{dom}(g \circ M)$. By definition of u and v, for $y \in \mathcal{X}$,

$$\begin{cases} f(y) - f(x) \geq \langle u, y - x \rangle \\ g(My) - g(Mx) \geq \langle v, M(y - x) \rangle = \langle M^* v, y - x \rangle . \end{cases}$$

Adding the two inequalities,

$$(f + g \circ M)(y) - (f + g \circ M)(x) \geq \langle \phi, y - x \rangle .$$

Thus, $\phi \in \partial(f + g \circ M)(x)$ and $\partial f(x) + M^* \partial g(Mx) \subset \partial(f + g \circ M)(x)$.

We will only prove the converse inclusion in the case where g is differentiable. Remark that as g is differentiable, $\operatorname{dom} g = \mathcal{Y}$ and thus the condition is trivially satisfied.
The general proof is omitted. It makes use of the Fenchel-Young inequality and of the strong duality theorem that will be given in Chapters 4 and 5.
So, suppose that g is differentiable and let $\phi \in \partial(f + g \circ M)(x)$. By definition of the subdifferential, for all $y \in \mathcal{X}$, $f(y) + g(My) \geq f(x) + g(Mx) + \langle \phi, y - x \rangle$. In particular, for all $h \in [0, 1]$,

$$f(x + h(y - x)) + g(M(x + h(y - x))) \geq f(x) + g(Mx) + h\langle \phi, y - x \rangle.$$

As f is convex, $f(x + h(y - x)) \leq hf(y) + (1 - h)f(x)$ and so, dividing by h,

$$f(y) \geq f(x) - \frac{1}{h}\big(g(M(x + h(y - x))) - g(x)\big) + \langle \phi, y - x \rangle.$$

We let h tend to 0 and we obtain $f(y) \geq f(x) + \langle -\nabla(g \circ M)(x) + \phi, y - x \rangle$. Said otherwise, $q - \nabla(g \circ M)(x) \in \partial f(x)$. We conclude using the chain rule and Proposition 2.3.2. □

2.5 Fermat's rule, optimality conditions.

A point x is called a **minimizer** of f if $f(x) \leq f(y)$ for all $y \in \mathcal{X}$. The set of minimizers of f is denoted $\arg\min(f)$.

Theorem 2.5.1 (Fermat's rule)**.** $x \in \arg\min f \;\Leftrightarrow\; 0 \in \partial f(x)$.

Proof. $x \in \arg\min f \;\Leftrightarrow\; \forall y, f(y) \geq f(x) + \langle 0, y - x\rangle \Leftrightarrow 0 \in \partial f(x)$. □

Recall that, in the differentiable, non convex case, a *necessary* condition (not a sufficient one) for $\bar{x}$ to be a local minimizer of f, is that $\nabla f(\bar{x}) = 0$. Convexity allows handling non differentiable functions, and turns the necessary condition into a sufficient one.
Besides, local minima for any function f are not necessarily global ones. In the convex case, everything works fine:

Proposition 2.5.1. *Let x be a local minimum of a convex function f. Then, x is a global minimizer.*

Proof. The local minimality assumption means that there exists an open ball $V \subset \mathcal{X}$, such that $x \in V$ and that, for all $u \in V$, $f(x) \leq f(u)$.
Let $y \in \mathcal{X}$ and t such that $u = x + t(y - x) \in V$. Then using the convexity of f, we get $f(u) \leq tf(y) + (1 - t)f(x)$. Re-organizing, we get

$$f(y) \geq t^{-1}\big(f(u) - (1 - t)f(x)\big) \geq f(x).$$

□

Exercise 2.5.1. Let us denote

$$\mathrm{prox}_g(y) = \arg\min_{x \in \mathcal{X}} g(x) + \frac{1}{2}\|y - x\|^2$$

the proximal operator of g at y.
Fix $\gamma > 0$. Show that the fixed points of the nonlinear equation

$$x = \mathrm{prox}_{\gamma g}(x - \gamma \nabla f(x))$$

are the minimizers of the function $F = f + g$.

Chapter 3

Primal methods

3.1 Gradient method

3.1.1 Constant step sizes

The gradient method is the most basic optimization method for a differentiable function f. It consists in a sequence $(x_k)_{k\in\mathbb{N}}$ of points in $\mathbb{R}^n$ defined by induction from $x_0 \in \mathbb{R}^n$ by

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k)$$

where for all k, γ_k is a positive coefficient.

Theorem 3.1.1. *Let f be a convex differentiable function that has a minimizer x^* and whose gradient is L-Lipschitz continuous. The gradient method with constant step size $\gamma_k = \frac{1}{L}$ satisfies*

$$f(x_k) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k}$$

If moreover f is μ-strongly convex, then

$$f(x_k) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f(x^*) + \tfrac{L}{2}\|x_0 - x^*\|^2)$$
$$\|x_k - x^*\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k (\tfrac{2}{L}(f(x_0) - f(x^*)) + \|x_0 - x^*\|^2)$$

Proof. We will prove a more general result in Theorem 3.3.1. □

3.1.2 Line search

Considering constant step sizes makes the proof easier but has drawbacks in practice:

- One needs to compute the Lipschitz constant of the gradient of f, which may be a non-negligible amount of work.
- Some functions, like $(x \mapsto x^4)$ simply do not have a Lipschitz gradient. However, the gradient may be locally Lipschitz.
- Even if the function has a Lipschitz gradient; the estimation of the Lipschitz constant may take into account regions where the curvature is large but that are never visited by the algorithm.

A solution to these three issues is a line search procedure. The idea of line search is to choose γ_k adaptively using local information.

Exact line search. We take

$$\gamma_k = \arg\min_{\gamma \in \mathbb{R}_+} f(x_k - \gamma \nabla f(x_k)).$$

This method is most efficient when we have a closed formula for the 1-dimensional oprimization problem.

Taylor-based line search. In the proof of convergence of Theorem 3.3.1, we only need the Lipschitz continuity of ∇f in order to ensure that

$$f(x_{k+1}) = f(x_k - \frac{1}{L}\nabla f(x_k)) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_k - x_{k+1}\|^2.$$

The Taylor-based line search chooses a step size γ_k such that for the tentative update defined by $x^+(\gamma_k) = x_k - \gamma_k \nabla f(x_k)$, we have

$$f(x^+(\gamma_k)) \leq f(x_k) + \langle \nabla f(x_k), x^+(\gamma_k) - x_k \rangle + \frac{1}{2\gamma_k}\|x_k - x^+(\gamma_k)\|^2$$

using the following algorithm.
We set $b > 0, a \in (0,1)$ and we find the first nonnegative integer l such that

$$f(x^+(ba^l)) \leq f(x_k) + \langle \nabla f(x_k), x^+(ba^l) - x_k \rangle + \frac{1}{2ba^l}\|x_k - x^+(ba^l)\|^2 \tag{3.1.1}$$

Then we set $\gamma_k = ba^l$. Clearly, if such an l exists, the desired inequality will hold.

Proposition 3.1.1. *If f has a L-Lipschitz gradient, then the Taylor-based line search will terminate with $l < \frac{log(ba^{-1}L)}{\log(a^{-1})}$ and $ba^l < aL$*

Proof. If ∇f is L-Lipschitz, then it is also L'-Lipschitz for all $L' \geq L$. Hence, as soon as $1/(ba^l) \geq L$ (which will eventually happen since $1/a > 1$), (3.1.1) will hold and the line search will terminate. Just before, we had $1/(ba^{l-1}) < L$, so, $1/(ba^l) < a^{-1}L$. We get the bound on l by passing to the log. □

Note that we do not need to know this Lipschitz constant in order to run the line search.
Classical choices for the parameters are $a = 0.5$ and $b = 2\gamma_{k-1}$.

Armijo's line search. This line search is the most famous one. Given $a \in (0,1)$, $b > 0$ and $\beta \in (0,1)$, determine the first integer l such that

$$f(x^+(ba^l)) \leq f(x_k) + \beta \langle \nabla f(x_k), x^+(ba^l) - x_k \rangle$$

In the case of gradient descent, $\langle \nabla f(x_k), x^+(\gamma) - x_k \rangle = -\gamma \|\nabla f(x_k)\|^2$ so we can see that the Taylor-based line search is equivalent to an Armijo's line search with $\beta = 1/2$.

3.2 Proximal point method

When the function to minimize is a general, non-differentiable convex function, the gradient method cannot be used. One may use subgradients instead of the gradient but the resulting algorithm requires step sizes that decrease to 0 and is thus very slow.

An alternative approach is the so-called proximal point method. It is based on the proximal operator of the function f to be minimized defined by

$$\text{prox}_f(x) = \arg\min_{y\in\mathbb{R}^n} f(y) + \frac{1}{2}\|y-x\|^2.$$

Computing this proximal operator may seem as hard as the original problem because it is also an optimization problem. But thanks to the $\frac{1}{2}\|y-x\|^2$ term, this new optimization problem is strongly convex, which make it much more tractable than the original problem $\min_x f(x)$.
The proximal point method consists in successively iterating proximal operator computations while updating the proximal center x_k:

$$x_{k+1} = \text{prox}_{\gamma f}(x_k) = \arg\min_{y\in\mathbb{R}^n} f(y) + \frac{1}{2\gamma}\|y-x_k\|^2$$

The parameter γ features a tradeoff between easier steps and the number of steps available. When γ is small, computing the proximal operator involves a $\frac{1}{\gamma}$-strongly convex function, which is easier to minimize, but very different from the original f. When γ is large, the function $(x \mapsto f(x) + \frac{1}{2\gamma}\|y-x\|^2)$ is not so different from f and the iterates will converge quicker. The proof of convergence of the proximal point method is given in a more general form in Theorem 3.3.1.

3.3 Proximal gradient method

The proximal gradient method is a method designed to solve composite problems of the type

$$\min_{x\in\mathbb{R}^n} f(x) + g(x)$$

where f has a Lipschitz gradient and the proximal operator of g is easy to compute, ideally, prox_g has a close form. The algorithm is given by

$$x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$$

for given step sizes $\gamma_k > 0$

Theorem 3.3.1. *Let f be a convex differentiable function whose gradient is L-Lipschitz continuous, g be a convex l.s.c. function and x^* a minimizer of $f+g$. The proximal gradient method with constant step size $\gamma_k = \frac{1}{L}$ satisfies*

$$f(x_k) + g(x_k) - f(x^*) - g(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k}$$

If moreover f is μ-strongly convex, then denoting $\Delta_0 = f(x_0)+g(x_0)-f(x^*)-g(x^*)+\frac{L}{2}\|x_0-x^*\|^2$,

$$f(x_k) + g(x_k) - f(x^*) - g(x^*) \leq \left(1-\frac{\mu}{L}\right)^k \Delta_0$$
$$\|x_k - x^*\|^2 \leq \left(1-\frac{\mu}{L}\right)^k \frac{2\Delta_0}{L}$$

Proof. *Cf.* tutorial. □

Remark 3.3.1. A slightly better rate can be obtained by taking $\gamma = \frac{2}{L+\mu}$, provided μ is known Nesterov (2004).

The Taylor-based line search can be generalized to the proximal gradient method. We need to choose γ_k such that for $x^+(\gamma_k) = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$, we have

$$f(x^+(\gamma_k)) \leq f(x_k) + \langle \nabla f(x_k), x^+(\gamma_k) - x_k \rangle + \frac{1}{2\gamma_k} \|x_k - x^+(\gamma_k)\|^2. \tag{3.3.1}$$

Armijo's line search can be generalized in the same way.

3.4 Newton's method

Newton's method uses the Hessian matrix in order to ensure convergence in a smaller number of iterations. As shown in MDI210 – Optimisation et analyse numérique we have a quadratic convergence.

Theorem 3.4.1. *If f is three times continuously differentiable and if x_0 is chosen close enough to a local minimum x^* where the Hessian matrix of f is positive definite, then the sequence x_k generated by Newton's method*

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

converges to x^* and there exists $M > 0$ such that

$$\|x_{k+1} - x^*\| \leq M \|x_k - x^*\|^2.$$

Out of the region of quadratic convergence, Newton's method may diverge. Hence, the method should be combined with a line search procedure similar to (3.3.1) in order to ensure convergence even if x_0 is not close to x^*.

Chapter 4

Fenchel-Legendre transform, dual problem

We introduce now an important tool of convex analysis, especially useful for duality approaches: the Fenchel-Legendre transform.

4.1 Fenchel-Legendre Conjugate

Definition 4.1.1. Let $f : \mathcal{X} \to [-\infty, +\infty]$. The **Fenchel-Legendre conjugate** of f is the function $f^* : \mathcal{X} \to [-\infty, \infty]$, defined by

$$f^*(\phi) = \sup_{x \in \mathcal{X}} \langle \phi, x \rangle - f(x)\,, \qquad \forall \phi \in \mathcal{X}.$$

Notice that

$$f^*(0) = -\inf_{x \in \mathcal{X}} f(x).$$

Figure 4.1 provides a graphical representation of f^*. You should get the intuition that, in the differentiable case, if the maximum is attained in the definition of f^* at point x_0, then $\phi = \nabla f(x_0)$, and $f^*(\phi) = \langle \nabla f(x_0), x_0 \rangle - f(x_0)$. This intuition will be proved correct in proposition 4.1.2.

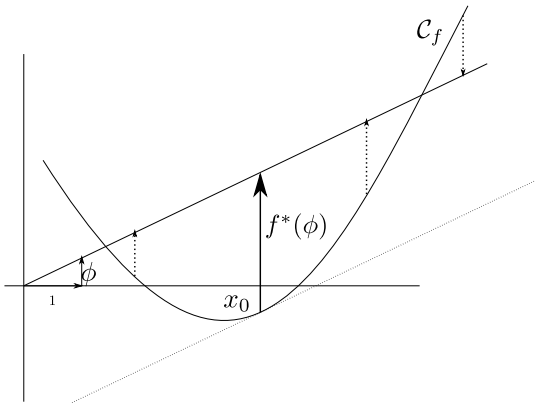

Figure 4.1: Fenchel Legendre transform of a smooth function f. The maximum positive difference between the line with slope $\tan(\phi)$ and the graph $\mathcal{C}_f$ of f is reached at x_0.

Exercise 4.1.1.
Prove the following statements.
General hint: If $h_\phi : x \mapsto \langle \phi, x\rangle - f(x)$ reaches a maximum at x^*, then $f^*(\phi) = h_\phi(x^*)$. Furthermore, h_ϕ is concave (if f is convex). If h_ϕ is differentiable, it is enough to find a zero of its gradient to obtain a maximum.
Indeed, $x \in \arg\min(-h_\phi) \Leftrightarrow 0 \in \partial(-h_\phi)$, and, if $-h_\phi$ is differentiable, $\partial(-h_\phi) = \{-\nabla h_\phi\}$.

1. If $\mathcal{X} = \mathbb{R}$ and f is a quadratic function (of the kind $f(x) = (x-a)^2 + b$), then f^* is also quadratic.

2. In $\mathbb{R}^n$, let A by a symmetric, definite positive matrix and $f(x) = \langle x, Ax\rangle$ (a quadratic function). Show that f^* is also quadratic.

3. $f : \mathcal{X} \to [-\infty, +\infty]$. Show that $f = f^* \Leftrightarrow f(x) = \frac{1}{2}\|x\|^2$.

   Hint: For the 'if' part: show first that $f(\phi) \geq \langle \phi, \phi\rangle - f(\phi)$.

   Then, show that $f(\phi) \leq \sup_x \langle \phi, x\rangle - \frac{1}{2}\|x\|^2$. Conclude.

4. $\mathcal{X} = \mathbb{R}$, for
$$f(x) = \begin{cases} 1/x & \text{if } x > 0; \\ +\infty & \text{otherwise .} \end{cases}$$
   we have,
$$f^*(\phi) = \begin{cases} -2\sqrt{-\phi} & \text{if } \phi \leq 0; \\ +\infty & \text{otherwise .} \end{cases}$$

5. $\mathcal{X} = \mathbb{R}$, if $f(x) = \exp(x)$, then
$$f^*(\phi) = \begin{cases} \phi\ln(\phi) - \phi & \text{if } \phi > 0; \\ 0 & \text{if } \phi = 0; \\ +\infty & \text{if } \phi < 0. \end{cases}$$

Notice that, if $f(x) = -\infty$ for some x, then $f^* \equiv +\infty$.
Nonetheless, under 'reasonable' conditions on f, the Legendre transform enjoys nice properties, and even f can be recovered from f^* (through the equality $f = f^{**}$, see Proposition 4.1.4).

Proposition 4.1.1 (Properties of f^*)**.**
Let $f : \mathcal{X} \to [-\infty, +\infty]$ be any function.

1. *f^* is always convex, and l.s.c.*

2. *If $\operatorname{dom} f \neq \emptyset$, then $-\infty \notin f^*(\mathcal{X})$*

3. *If f is convex and proper, then f^* is convex, l.s.c., proper.*

Proof.

1. Fix $x \in \mathcal{X}$ and consider the function $h_x : \phi \mapsto \langle \phi, x\rangle - f(x)$. From the definition, $f^* = \sup_{x\in\mathcal{X}} h_x$. Each h_x is affine, whence convex. Using proposition 2.1.2, f^* is also convex. Furthermore, each h_x is continuous, whence l.s.c, so that its epigraph is closed. Lemma 2.2.3 thus shows that f^* is l.s.c.

2. From the hypothesis, there is an x_0 in $\operatorname{dom} f$. Let $\phi \in \mathcal{X}$. The result is immediate:

$$f^*(\phi) \geq h_{x_0}(\phi) = f(x_0) - \langle \phi, x_0 \rangle > -\infty.$$

3. In view of points 1. and 2., it only remains to show that $f^* \not\equiv +\infty$. Let $x_0 \in \operatorname{relint}(\operatorname{dom} f)$. According to proposition 2.3.3, there exists a subgradient ϕ_0 of f at x_0. Moreover, since f is proper, $f(x_0) < \infty$. From the definition of a subgradient,

$$\forall x \in \operatorname{dom} f, \langle \phi_0, x - x_0 \rangle \leq f(x) - f(x_0).$$

Whence, for all $x \in \mathcal{X}$,

$$\langle \phi_0, x \rangle - f(x) \leq \langle \phi_0, x_0 \rangle - f(x_0),$$

thus, $\sup_x \langle \phi_0, x \rangle - f(x) \leq \langle \phi_0, x_0 \rangle - f(x_0) < +\infty$.

Therefore, $f^*(\phi_0) < +\infty$. □

Proposition 4.1.2 (Fenchel-Young)**.** *Let $f : \mathcal{X} \to [-\infty, \infty]$. For all $(x, \phi) \in \mathcal{X}^2$, the following inequality holds:*

$$f(x) + f^*(\phi) \geq \langle \phi, x \rangle ,$$

With equality if and only if $\phi \in \partial f(x)$.

Proof. The inequality is an immediate consequence of the definition of f^*. The condition for equality to hold (*i.e.*, for the converse inequality to be valid), is obtained with the equivalence

$$f(x) + f^*(\phi) \leq \langle \phi, x \rangle \;\Leftrightarrow\; \forall y,\; f(x) + \langle \phi, y \rangle - f(y) \leq \langle \phi, x \rangle \;\Leftrightarrow\; \phi \in \partial f(x) .$$

□

Proposition 4.1.3. *Let $f^{**} = (f^*)^*$ be the biconjgate of f. For all $x \in \mathcal{X}$,*

$$f(x) \geq f^{**}(x)$$

Proof. The Fenchel-Young inequality writes as: for all $\phi \in \mathcal{X}$,

$$f(x) \geq \langle \phi, x \rangle - f^*(\phi).$$

Taking the supremum on ϕ in the right hand side of the inequality gives the result. □

Proposition 4.1.4 (Involution property, Fenchel-Moreau)**.** *If f is convex, l.s.c. and proper, then $f = f^{**}$.*

Proof. Admitted. The inequality $f^{**} \leq f$ is less trivial. It is based on the fact that a convex, l.s.c. and proper function is equal to its affine minorants. □

Exercise 4.1.2. Let $f : \mathcal{X} \to (-\infty, +\infty]$ a proper, convex, l.s.c. function. Show that

$$\partial(f^*) = (\partial f)^{-1}$$

where, for $\phi \in \mathcal{X}$, $(\partial f)^{-1}(\phi) = \{x \in \mathcal{X} :\ \phi \in \partial f(x)\}$.
Hint: Use Fenchel-Young inequality to show one inclusion, and the property $f = f^{**}$ for the other one.

4.2 Lagrangian function

In this chapter, we consider the convex optimization problem

$$\text{minimize over } \mathbb{R}^n : \quad f(x) + \iota_{g \preceq 0}(x) + \iota_{A=0}(x) \ . \tag{4.2.1}$$

(*i.e.* minimize $f(x)$ over $\mathbb{R}^n$, under the constraint $g(x) \preceq 0$ and $A(x) = 0$),
where $f : \mathbb{R}^n \to (-\infty, \infty]$ is convex and proper; $g(x) = (g_1(x), \dots, g_p(x))$, each $g_i : \mathbb{R}^n \to \mathbb{R}$ is a convex function $(1 \leq i \leq p)$; $A : \mathbb{R}^n \to \mathbb{R}^m$ is an affine function and $\iota_{g \preceq 0} = \iota_{g^{-1}(\mathbb{R}^p_-)}$.

Under these conditions, the function $x \mapsto f(x) + \iota_{g \preceq 0}(x) + \iota_{A=0}(x)$ is convex.

Definition 4.2.1 (primal value, primal optimal point)**.** The **primal value** associated to (4.2.1) is the infimum

$$p = \inf_{x \in \mathbb{R}^n} f(x) + \iota_{g \preceq 0}(x) + \iota_{A=0}(x).$$

A point $x^* \in \mathbb{R}^n$ is called **primal optimal** if

$$p = f(x^*) + \iota_{g \preceq 0}(x^*) + \iota_{A=0}(x^*).$$

Notice that, under our assumption, $p \in [-\infty, \infty]$. Also, there is no guarantee about the existence of a primal optimal point, *i.e.* that the primal value be attained.

Since (4.2.2) may be difficult to solve, it is useful to see this as an 'inf sup' problem, and solve a 'sup inf' problem instead (see definition 4.3.1 below). To make this precise, we introduce the Lagrangian function.

Definition 4.2.2. The **Lagrangian function** associated to problem (4.2.1) is the function

$$\begin{aligned} L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p &\longrightarrow [-\infty, +\infty] \\ (x, \phi_E, \phi_I) &\mapsto f(x) + \langle \phi_E, A(x) \rangle + \langle \phi_I, g(x) \rangle - I_{\mathbb{R}^p_+}(\phi_I) \end{aligned}$$

(where $\mathbb{R}^p_+ = \{\phi \in \mathbb{R}^p, \phi \succeq 0\}$).

The link with the initial problem comes next:

Lemma 4.2.1 (constrained objective as a supremum)**.** *The constrained objective is the supremum (over $\phi = (\phi_E, \phi_I)$) of the Lagrangian function,*

$$\forall x \in \mathbb{R}^n, \ f(x) + \iota_{g \preceq 0}(x) + \iota_{A=0}(x) = \sup_{\phi \in \mathbb{R}^{m+p}} L(x, \phi)$$

Proof. We give the proof in the case $m = 0$ (only inequality constraints). The general case is similar.
Distinguish the cases $g(x) \preceq 0$ and $g(x) \not\preceq 0$.
(a) If $g(x) \not\preceq 0, \exists i \in \{1, \dots, p\} : g_i(x) > 0$. Choosing $\phi_t = te_i$ (where $\mathbf{e} = (e_1, \dots, e_p)$ is the canonical basis of $\mathbb{R}^p$)), $t \geq 0$, then $\lim_{t \to \infty} L(x, \phi_t) = +\infty$, whence $\sup_{\phi \in \mathbb{R}^p_+} L(x, \phi) = +\infty$. On the other hand, in such a case, $\iota_{g \preceq 0}(x) = +\infty$, whence the result.

(b) If $g(x) \leq 0$, then $\forall \phi \in \mathbb{R}^p_+, \langle \phi, g(x) \rangle \leq 0$, and the supremum is attained at $\phi = 0$. Whence, $\sup_{\phi \geq 0} L(x, \phi) = \sup_{\phi \in \mathbb{R}^p} L(x, \phi) = f(x)$.
On the other hand, $\iota_{g \preceq 0}(x) = 0$, so $f(x) + \iota_{g \preceq 0}(x) = f(x)$. The result follows. □

Equipped with lemma 4.2.1, the primal value associated to problem (4.2.1) writes

$$p = \inf_{x \in \mathbb{R}^n} \sup_{\phi \in \mathbb{R}^{m+p}} L(x, \phi). \tag{4.2.2}$$

One natural idea is to exchange the order of inf and sup in the above problem. Before proceeding, the following simple lemma allows to understand the consequence of such an exchange.

Proposition 4.2.1. *Let $F : A \times B \to [-\infty, \infty]$ any function. Then,*

$$\sup_{y \in B} \inf_{x \in A} F(x, y) \leq \inf_{x \in A} \sup_{y \in B} F(x, y).$$

Proof. $\forall (\bar{x}, \bar{y}) \in A \times B$,

$$\inf_{x \in A} F(x, \bar{y}) \leq F(\bar{x}, \bar{y}) \leq \sup_{y \in B} F(\bar{x}, y).$$

Taking the supremum over $\bar{y}$ in the left-hand side we still have

$$\sup_{\bar{y} \in B} \inf_{x \in A} F(x, \bar{y}) \leq \sup_{y \in B} F(\bar{x}, y).$$

Now, taking the infimum over $\bar{x}$ in the right-hand side yields

$$\sup_{\bar{y} \in B} \inf_{x \in A} F(x, \bar{y}) \leq \inf_{\bar{x} \in A} \sup_{y \in B} F(\bar{x}, y).$$

up to a simple change of notation, this is the expected result. □

4.3 Dual problem

Definition 4.3.1 (Dual problem, dual function, dual value)**.**
The **dual value** associated to (4.2.2) is

$$d = \sup_{\phi \in \mathbb{R}^{m+p}} \inf_{x \in \mathbb{R}^n} L(x, \phi_I, \phi_E).$$

The function

$$\mathcal{D}(\phi) = \inf_{x \in \mathbb{R}^n} L(x, \phi)$$

is called the **Lagrangian dual function**. Thus, the **dual problem** associated to the primal problem (4.2.1) is

$$\text{maximize over } \mathbb{R}^{m+p} : \quad \mathcal{D}(\phi).$$

A vector $\lambda \in \mathbb{R}_+ p$ is called **dual optimal** if

$$d = \mathcal{D}(\lambda).$$

Without any further assumption, there is no reason for the two values (primal and dual) to coincide. However, as a direct consequence of Proposition 4.2.1, we have :

Proposition 4.3.1 (Weak duality)**.** *Let p and d denote respectively the primal and dual value for problem (4.2.1). Then,*

$$d \leq p \ .$$

Proof. Apply Proposition 4.2.1. □

Definition 4.3.2 (Saddle point)**.** Let $F : A \times B \to [-\infty, \infty]$ any function, and A, B two sets. The point $(x^*, y^*) \in A \times B$ is called a **saddle point** of F if, for all $(x, y) \in A \times B$,

$$F(x^*, y) \leq F(x^*, y^*) \leq F(x, y^*).$$

Proposition 4.3.2. *Let $F : A \times B \to [-\infty, \infty]$. F has a saddle point (x^*, y^*) if and only if*

$$\sup_{y \in B} \inf_{x \in A} F(x, y) = \inf_{x \in A} F(x, y^*) = F(x^*, y^*) = \sup_{y \in B} F(x^*, y) = \inf_{x \in A} \sup_{y \in B} F(x, y).$$

Proof. Suppose F has a saddle point (x^*, y^*). As $\forall y$, $F(x^*, y) \leq F(x^*, y^*)$, we take the supremum in y to get $\sup_{y \in B} F(x^*, y) \leq F(x^*, y^*)$.

$$\begin{aligned}
\sup_{y \in B} \inf_{x \in A} F(x, y) &\leq \inf_{x \in A} \sup_{y \in B} F(x, y) && \text{(Prop. 4.2.1)} \\
&\leq \sup_{y \in B} F(x^*, y) && \text{(def of inf)} \\
&\leq F(x^*, y^*) && \text{(saddle point)} \\
&\leq \inf_{x \in A} F(x, y^*) && \text{(saddle point)} \\
&\leq \sup_{y \in B} \inf_{x \in A} F(x, y) && \text{(def of sup)}
\end{aligned}$$

Hence all inequalities are equalities and we get the result.
The converse implication is straightforward using the two inner equalities. □

Chapter 5

Strong duality theorem

5.1 Equality constraints

Theorem 5.1.1. *Let A be an affine function and f be a convex function. Let us consider the problem with equality constraints*

$$\min_{x\in\mathbb{R}^n} f(x) + \iota_{A=0}(x),$$

the associated Lagrangian

$$L(x,\phi) = f(x) + \langle \phi, A(x)\rangle$$

and the dual problem

$$\sup_{\phi\in\mathbb{R}^m} \mathcal{D}(\phi)$$

where $\mathcal{D}(\phi) = \inf_{x\in\mathbb{R}^n} f(x) + \langle \phi, A(x)\rangle$.
If $0 \in \operatorname{relint}(A(\operatorname{dom} f))$ *(constraint qualification condition), then*

1. $\inf_{x\in\mathbb{R}^n} f(x) + \iota_{A=0}(x) < +\infty$
2. $\inf_{x\in\mathbb{R}^n} f(x) + \iota_{A=0}(x) = \sup_{\phi\in\mathbb{R}^m} \mathcal{D}(\phi)$ (i.e., *the duality gap is zero).*
3. *(Dual attainment at some λ):*

$$\exists \lambda \in \mathbb{R}^m, \text{such that } \; d = \mathcal{D}(\lambda).$$

If moreover, $\exists x^* \in \mathbb{R}^n$ *such that* $\min_{x\in\mathbb{R}^n} f(x) + \iota_{A=0}(x) = f(x^*) + \iota_{A=0}(x^*)$, *then* (x^*, λ) *is a saddle point of* L.

Note that if $\operatorname{dom} f = \mathbb{R}^n$, then the constraints are automatically qualified.

Proof of the theorem, introduction. The idea of the proof is to apply Propositions 2.3.3 and 4.1.3 on the value function

$$\mathcal{V}(b) = \inf_{x\in\mathbb{R}^n} f(x) + \iota_{\{b\}}(A(x)).$$

Note that $\mathcal{V}(0)$ is the value of the primal problem. $\mathcal{V}$ is convex since it is the infimum of a jointly convex function (Proposition 2.1.3). To apply Proposition 4.1.3, we need to calculate $\mathcal{V}^{**}$ and thus $\mathcal{V}^*$.

Lemma 5.1.1. *For all* $\phi \in \mathbb{R}^m$, $\mathcal{V}^*(-\phi) = -\mathcal{D}(\phi)$.

Proof. For $\phi \in \mathbb{R}^m$, by definition of the Fenchel conjugate,

$$\begin{aligned}
\mathcal{V}^*(-\phi) &= \sup_{y\in\mathbb{R}^m} \langle -\phi, y\rangle - \mathcal{V}(y) \\
&= \sup_{y\in\mathbb{R}^m} \langle -\phi, y\rangle - \inf_{x\in\mathbb{R}^n} \big[f(x) + \iota_{\{y\}}(A(x))\big] \\
&= \sup_{y\in\mathbb{R}^m} \langle -\phi, y\rangle + \sup_{x\in\mathbb{R}^n} \big[-f(x) - \iota_{\{y\}}(A(x))\big] \\
&= \sup_{y\in\mathbb{R}^m} \sup_{x\in\mathbb{R}^n} \langle -\phi, y\rangle - f(x) - \iota_{\{y\}}(A(x)) \\
&= \sup_{x\in\mathbb{R}^n} \Big[\sup_{y\in\mathbb{R}^m} \underbrace{\langle -\phi, y\rangle - \iota_{\{y\}}(A(x))}_{\varphi_x(y)} \Big] - f(x). \qquad (5.1.1)
\end{aligned}$$

For a fixed $x \in \operatorname{dom} f$, consider the function $\varphi_x : y \mapsto \langle -\phi, y\rangle - \iota_{\{y\}}(A(x))$. As

$$\varphi(y) = \begin{cases} -\infty & \text{if } y \neq A(x) \\ \langle -\phi, A(x)\rangle & \text{otherwise,} \end{cases}$$

(5.1.1) becomes

$$\begin{aligned}
\mathcal{V}^*(-\phi) &= \sup_{x\in\mathbb{R}^n} \langle -\phi, A(x)\rangle - f(x) = -\inf_{x\in\mathbb{R}^n} \underbrace{f(x) + \langle \phi, A(x)\rangle}_{L(x,\phi)} \\
&= -\mathcal{D}(\phi)
\end{aligned}$$

□

Corollary 5.1.1. *The dual function $\mathcal{D}$ is concave and upper semi-continuous.*

Proof. Proposition 4.1.1. □

Proof of the theorem, continued. From Lemma 5.1.1, we deduce

$$\begin{aligned}
\mathcal{V}^{**}(0) &= \sup_{\phi\in\mathbb{R}^p} -\mathcal{V}^*(\phi) = \sup_{\phi\in\mathbb{R}^p} -\mathcal{V}^*(-\phi) \quad \text{(by symmetry of } \mathbb{R}^p) \\
&= \sup_{\phi\in\mathbb{R}^p} \mathcal{D}(\phi)
\end{aligned}$$

Hence, $\mathcal{V}^{**}(0)$ is the value of the dual problem. By Proposition 4.1.3, $\mathcal{V}(0) \geq \mathcal{V}^{**}(0)$. Said otherwise, $\inf_{x\in\mathbb{R}^n} f(x) + \iota_{A=0}(x) \geq \sup_{\phi\in\mathbb{R}^p} \mathcal{D}(\phi)$ and we recover weak duality.
Now remark that $\operatorname{dom} \mathcal{V} = \{b \in \mathbb{R}^m \ : \ \exists x \in \operatorname{dom} f, A(x) = b\} = A(\operatorname{dom} f)$. So the constraint qualification condition $0 \in \operatorname{relint}(A(\operatorname{dom} f))$ is equivalent to $0 \in \operatorname{relint}(\operatorname{dom} \mathcal{V})$ and we can apply Proposition 2.3.3: $\partial\mathcal{V}(0) \neq \emptyset$.
To show the dual attainment, we take $\lambda \in \partial\mathcal{V}(0) \neq \emptyset$. Equality in Fenchel-Young (Proposition 4.1.2) writes: $\mathcal{V}(0) + \mathcal{V}^*(\lambda) = \langle \lambda, 0\rangle = 0$. Thus, we have

$$\begin{aligned}
\mathcal{V}(0) &= -\mathcal{V}^*(\lambda) \\
&= \mathcal{D}(-\lambda) \\
&\leq \sup_{\phi} \mathcal{D}(\phi) = \mathcal{V}^{**}(0) \leq \mathcal{V}(0)
\end{aligned}$$

Hence, all the inequalities are equalities:

$$\inf_{x\in\mathbb{R}^n} f(x) + \iota_{A=0}(x) = \mathcal{V}(0) = -\mathcal{V}^*(\lambda) = \mathcal{D}(-\lambda) = \sup_{\phi} \mathcal{D}(\phi).$$

This shows that the duality gap is 0 and that $-\lambda$ is a dual optimum.
Now, if there exists a primal optimum x^*,

$$\mathcal{V}(0) = f(x^*) + \iota_{A=0}(x^*) = \sup_{\phi \in \mathbb{R}^m} L(x^*, \phi) \geq L(x^*, -\lambda) \geq \inf_x L(x, -\lambda) = \mathcal{D}(-\lambda)$$

and we conclude using the fact that $\mathcal{V}(0) = \mathcal{D}(-\lambda)$ and Proposition 4.3.2. □

Remark 5.1.1. The proof shows that the negative subgradients of the value function at 0 are optimal dual points. Hence, we can interpret the optimal dual points as the sensitivity of the primal objective to changes in the constraint.

Exercise 5.1.1. The goal of this exercise is to study the following semi-definite program where the variables are semi-definite matrices (the set of semi-definite matrices is denoted S_+^n)

$$\begin{aligned} &\inf_{x \in \mathbb{R}^{3\times 3}} X_{3,3} + \iota_{S_+^3}(X) \\ &\text{under the constraints: } X_{1,2} + X_{2,1} + X_{3,3} = 1 \\ &X_{2,2} = 0 \end{aligned}$$

We will denote $E_{3,3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $f(x) = \langle E_{3,3}, X \rangle + \iota_{S_+^3}(X)$, $G(X) = [X_{1,2} + X_{2,1} + X_{3,3}; X_{2,2}]$, $e_1 = [1; 0]$ and $A(X) = G(X) - e_1$.

1. Give an optimal solution to this problem.
   Hint: determine the set of feasible points.
2. Compute the dual of this problem and solve it. What do you observe?
3. What does $0 \in A(\operatorname{dom} f)$ mean for the feasibility of the optimization problem?
4. Show that $[0; \epsilon] \in A(\operatorname{dom} f)$ if and only if $\epsilon \geq 0$. Deduce that the constraints are not qualified.

Exercise 5.1.2. Find the dual of the Lasso problem.

Exercise 5.1.3. Find the dual of the Support Vector Machine problem.

5.2 Inequality constraints

Theorem 5.2.1. *Let us consider the problem with inequality constraints*

$$\min_{x \in \mathbb{R}^n} f(x) + \iota_{g \leq 0}(x),$$

the associated Lagrangian

$$L(x, \phi) = f(x) + \langle \phi, g(x) \rangle - \iota_{\mathbb{R}_+^p}(\phi)$$

and the dual problem

$$\sup_{\phi \in \mathbb{R}^p} \mathcal{D}(\phi)$$

where $\mathcal{D}(\phi) = \inf_{x \in \mathbb{R}^n} f(x) + \langle \phi, g(x) \rangle - \iota_{\mathbb{R}_+^p}(\phi)$.
If $\exists x_0 \in \operatorname{dom} f$ *such that for all* j, $g_j(x_0) < 0$ *(Slater's constraint qualification condition), then*

1. $\inf_{x\in\mathbb{R}^n} f(x) + \iota_{g\leq 0}(x) < +\infty$
2. $\inf_{x\in\mathbb{R}^n} f(x) + \iota_{g\leq 0}(x) = \sup_{\phi\in\mathbb{R}^p} \mathcal{D}(\phi)$ (i.e., *the duality gap is zero).*
3. *(Dual attainment at some λ):*

$$\exists \lambda \in \mathbb{R}^p_+, \textit{such that } \; d = \mathcal{D}(\lambda).$$

If moreover, $\exists x^* \in \mathbb{R}^n$ such that $\min_{x\in\mathbb{R}^n} f(x) + \iota_{g\leq 0}(x) = f(x^*) + \iota_{g\leq 0}(x^*)$, then (x^*, λ) is a saddle point of L.

Proof. The proof is similar to the equality case. The main difference is in the domain of the value function:

$$\operatorname{dom} \mathcal{V} = \{b \in \mathbb{R}^p \; : \; \exists x \in \operatorname{dom} f, g(x) \leq b\}.$$

Slater's condition is exactly saying that $0 \in \operatorname{int} \operatorname{dom} \mathcal{V}$. □

Remark 5.2.1. The Karush-Kuhn-Tucker conditions can be recovered by writing Fermat's rule for the inf-sup conditions and the fact that $\operatorname{dom} g_j = \mathbb{R}^n$ for all j:

$$\begin{aligned} 0 \in \partial_x L(x^*, \phi^*) &= \sum_{j=1}^{p} \phi_j^* \partial g_j(x^*) + \partial f(x^*) \\ 0 \in \partial_\phi(-L)(x^*, \phi^*) &= -g(x^*) + \partial \iota_{\mathbb{R}^p_+}(\phi^*). \end{aligned}$$

For this second condition, we need to compute $\partial \iota_{\mathbb{R}^p_+}$. First note that for all $\phi \in \mathbb{R}^p$, $\iota_{\mathbb{R}^p_+}(\phi) = \sum_{j=1}^p \iota_{\mathbb{R}_+}(\phi_j)$ and so $\partial \iota_{\mathbb{R}^p_+}(\phi) = \partial \iota_{\mathbb{R}_+}(\phi_1) \times \ldots \times \partial \iota_{\mathbb{R}_+}(\phi_n)$.
It is a good exercise to show that

$$\partial \iota_{\mathbb{R}_+}(\phi_j) = \begin{cases} \{0\} & \text{if } \phi_j > 0, \\ \mathbb{R}_- & \text{if } \phi_j = 0, \\ \emptyset & \text{if } \phi_j < 0. \end{cases}$$

We obtain that $\forall j$, $g_j(x^*) = 0$ if $\phi_j^* > 0$, $g_j(x^*) \leq 0$ if $\phi_j^* = 0$ and that ϕ_j^* is never strictly negative.
To sum up, at a saddle point (x^*, ϕ^*), we have

$$\begin{aligned} &0 \in \sum_{j=1}^{p} \phi_j^* \partial g_j(x^*) + \partial f(x^*) \\ &g(x^*) \leq 0\,, \qquad \phi^* \geq 0\,, \qquad \langle \phi^*, g(x^*) \rangle = 0 \end{aligned}$$

Exercise 5.2.1 (Examples of duals, Borwein and Lewis (2006), chap.4)**.** Compute the dual of the following problems. In other words, calculate the dual function $\mathcal{D}$ and write the problem of maximizing the latter as a convex minimization problem.

1. Linear program

$$\inf_{x\in\mathbb{R}^n} \langle c, x \rangle$$
$$\text{under constraint } Gx \preceq b$$

   where $c \in \mathbb{R}^n$, $b \in \mathbb{R}^p$ and $G \in \mathbb{R}^{p\times n}$.

   Hint: you should find that the dual problem is again a linear program, with equality constraints.

2. Quadratic program

$$\inf_{x\in\mathbb{R}^n} \frac{1}{2}\langle x, Cx\rangle$$
$$\text{under constraint } Gx \preceq b$$

where C is symmetric, positive, definite.

Hint : you should obtain an unconstrained quadratic problem.

- Assume in addition that the constraints are linearly independent, *i.e.* $\text{rank}(G) = p$, *i.e.* $G = \begin{pmatrix} w_1^\top \\ \vdots \\ w_p^\top \end{pmatrix}$, where $(w_1, \dots, w_p)$ are linearly independent. Compute then the dual value.

Exercise 5.2.2 (dual gap)**.** Consider the two examples in exercise 5.2.1, and assume, as in example 2., that the constraints are linearly independent. Show that the duality gap is zero under this conditions.
Hint: Slater.

5.3 Examples, Exercises and Problems

In addition to the following exercises, a large number of feasible and instructive exercises can be found in Boyd and Vandenberghe (2009), chapter 5, pp 273-287.

Exercise 5.3.1 (Max-entropy)**.** Let $p = (p_1, \dots, p_n)$, $p_i > 0$, $\sum_i p_i = 1$ a probability distribution over a finite set. If $x = (x_1, \dots, x_n)$ is another probability distribution ($x_i \geq 0$), and if we use the convention $0 \log 0 = 0$, the entropy of x with respect to p is

$$H_p(x) = -\sum_{i=1}^n x_i \log \frac{x_i}{p_i}.$$

To deal with the case $x_i < 0$, introduce the function $\psi : \mathbb{R} \to (-\infty, \infty]$:

$$\psi(u) = \begin{cases} u\log(u) & \text{if } u > 0 \\ 0 & \text{if } u = 0 \\ +\infty & \text{otherwise .} \end{cases}$$

If $g : \mathbb{R}^n \to \mathbb{R}^p$, the general formulation of the max-entropy problem under constraint $g(x) \preceq 0$ is

$$\texttt{maximize}_{\text{ over } \mathbb{R}^n} \quad \sum_i \big(-\psi(x_i) + x_i \log(p_i)\big)$$
$$\text{under constraints } \sum x_i = 1; g(x) \preceq 0.$$

In terms of minimization, the problem writes

$$\inf_{x\in\mathbb{R}^n} \sum_{i=1}^n \psi(x_i) - \langle x, c\rangle + \iota_{\langle \mathbf{1}_n, \cdot\rangle = 1}(x) + \iota_{g\preceq 0}(x). \tag{5.3.1}$$

with $c = \log(p) = (\log(p_1), \dots, \log(p_n))$ and $\mathbf{1}_n = (1, \dots, 1)$ (the vector of size n which coordinates are equal to 1).

A: preliminary questions

1. Show that
$$\partial \iota_{\langle \mathbf{1}_n, \cdot \rangle = 1}(x) = \begin{cases} \{\lambda_0 \mathbf{1}_n \ : \ \lambda_0 \in \mathbb{R}\} := \mathbb{R}\mathbf{1}_n & \text{if } \sum_i x_i = 1 \\ \emptyset & \text{otherwise.} \end{cases}$$

2. Show that ψ is convex
hint: compute first the Fenchel conjugate of the function exp, then use proposition 4.1.1.
Compute $\partial\psi(u)$ for $u \in \mathbb{R}$.

3. Show that
$$\partial(\sum_i \psi(x_i)) = \begin{cases} \sum_i (\log(x_i) + 1)\, \mathbf{e}_i & \text{if } x \succ 0 \\ \emptyset & \text{otherwise,} \end{cases}$$
where $(\mathbf{e}_i, \dots, \mathbf{e}_n)$ is the canonical basis of $\mathbb{R}^n$.

4. Check that, for any set A, $A + \emptyset = \emptyset$.

5. Consider the unconstrained optimization problem, (5.3.1) where the term $\iota_{g(x) \preceq 0}$ has been removed. Show that there exists a unique primal optimal solution, which is $x^* = p$.
Hint: Do not use Lagrange duality, apply Fermat's rule (section 2.5) instead. Then, check that the conditions for subdifferential calculus rules (proposition 2.4.1) apply.

B: Linear inequality constraints In the sequel, we assume that the constraints are linear, independent, and independent from $\mathbf{1}_n$, *i.e.*: $g(x) = Gx - b$, where $b \in \mathbb{R}^p$, and G is a $p \times n$ matrix,
$$G = \begin{pmatrix} (\mathbf{w}^1)^\top \\ \vdots \\ (\mathbf{w}^p)^\top \end{pmatrix},$$
where $\mathbf{w}^j \in \mathbb{R}^n$, and the vectors $(\mathbf{w}^1, \dots, \mathbf{w}^p, \mathbf{1}_n)$ are linearly independent. We also assume the existence of some point $\hat{x} \in \mathbb{R}^n$, such that
$$\forall i, \hat{x}_i > 0, \ \sum_i \hat{x}_i = 1, \ G\hat{x} = b. \tag{5.3.2}$$

1. Show that the constraints are qualified.
Hint: Show that $0 \in \mathrm{int}(G(\Sigma_n) - \{b\} + \mathbb{R}^p_+)$, where $\Sigma_n = \{x \in \mathbb{R}^n \ : \ x \succeq 0, \sum_i x_i = 1\}$ is the n-dimensional simplex. In other words, show that for all $y \in \mathbb{R}^p$ close enough to 0, there is some small $\bar{u} \in \mathbb{R}^n$, such that $x = \hat{x} + \bar{u}$ is admissible for problem (5.3.1), and $G\,x \leq b + y$.
To do so, exhibit some $u \in \mathbb{R}^n$ such that $Gu = -\mathbf{1}_p$ and $\sum u_i = 0$ (why does it exist?) Pick t such that $\hat{x} + tu > 0$. Finally, consider the 'threshold' $Y = -t\mathbf{1}_p \prec 0$ and show that, if $y \succ Y$, then $G(\hat{x} + tu) \leq b + y$. Conclude.

2. Using the KKT conditions, show that any primal optimal point x^* must satisfy:
$\exists Z > 0, \exists \lambda \in \mathbb{R}^p_+$:
$$x^*_i = \frac{1}{Z}\, p_i \ \exp\big[- \sum_{j=1}^{p} \lambda_j \mathbf{w}^j_i \big] \quad (i \in \{1, \dots, n\})$$
(this is a Gibbs-type distribution).

Chapter 6

Dual methods

6.1 Lagrange multipliers Method

6.1.1 Problem setting

In this section, we seek to solve the following problem:

$$\min_{x:Ax=0} f(x) \tag{6.1.1}$$

where $f : \mathcal{X} \to (-\infty, +\infty]$ is proper closed convex function, A is a linear operator on $\mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean sets. The Lagrangian function associated with the above problem is

$$\mathcal{L}(x, \lambda) = f(x) + \langle \lambda, Ax \rangle$$

and the corresponding dual function is given by

$$\begin{aligned} \Phi(\lambda) &= \inf_{x \in \mathcal{X}} \mathcal{L}(x, \lambda) \\ &= -\sup_{x \in \mathcal{X}} \left\langle -A^T\lambda, x \right\rangle - f(x) \\ &= -f^*(-A^T\lambda)\,. \end{aligned}$$

Therefore, the dual problem boils down to:

$$\min_{\lambda \in \mathcal{Y}} f^*(-A^T\lambda)\,.$$

In the sequel, we provide two methods for solving this dual problem.

6.1.2 Algorithm

In this paragraph, we restrict our attention to the special case where f is μ-strongly convex. This assumption can be quite restrictive in practice, and we shall see later some alternative methods that can be used in a broader setting. We start with the following lemma.

Lemma 6.1.1. *If f is μ-strongly convex, then f^* is differentiable and ∇f^* is μ^{-1}-Lipschitz continuous.*

Proof. Let $\lambda \in \mathcal{Y}$. By the Fenchel-Young property 4.1.2, $x \in \partial f^*(\lambda)$ if and only if $\lambda \in \partial f(x)$. By Fermat's rule, this is again equivalent to $x \in \arg\min f - \langle \lambda, \,.\, \rangle$. As f is strongly convex, the argument of the minimum exists and is unique. As such, x is well and uniquely defined. Thus, $\partial f^*(\lambda)$ is a singleton. Otherwise stated, f^* is differentiable.

We verify the Lipschitz continuity of ∇f^*. Let us fix λ and λ'. Set $x = \nabla f^*(\lambda)$ and $y = \nabla f^*(\lambda')$. Since $\lambda \in \partial f(x)$, the strong convexity of f implies

$$f(y) \geq f(x) + \langle \lambda, y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

and a similar inequality hold by symetry:

$$f(x) \geq f(y) + \langle \lambda', x - y \rangle + \frac{\mu}{2}\|y - x\|^2$$

Summing these inequalities leads to $0 \geq \langle \lambda - \lambda', y - x \rangle + \mu\|x - y\|^2$. Hence, by the Cauchy-Schwarz inequality, we have $\|y - x\|^2 \leq \frac{1}{\mu}\|\lambda - \lambda'\| \, \|x - y\|$. Thus $\|y - x\| \leq \frac{1}{\mu}\|\lambda - \lambda'\|$ and the proof is complete. □

Remark 6.1.1. Lemma 6.1.1 has a converse. We refer to (Hiriart-Urruty and Lemaréchal, 2012, Theorem 4.2.2).

Therefore, is f is μ-strongly convex, the (negative) dual function $\lambda \mapsto f^*(-A^T\lambda)$ is differentiable and its gradient is as well Lipschitz continuous. In these circumstances, the previous chapter indicates that a gradient descent method can be used in order to solve the dual problem. The gradient descent writes:

$$\begin{aligned}\lambda^{k+1} &= \lambda^k - \gamma \nabla (f^* \circ (-A^T))(\lambda^k) \\ &= \lambda^k + \gamma A \nabla f^*(-A^T\lambda^k)\end{aligned}$$

where $\gamma > 0$ is the step size of the gradient descent. Define $x^{k+1} = \nabla f^*(-A^T\lambda^k)$. By the Fenchel-Young property 4.1.2, this is equivalent to $-A^T\lambda^k \in \partial f(x^{k+1})$ or equivalently

$$0 \in \partial f(x^{k+1}) + A^T\lambda^k \, .$$

By Fermat's rule, the above inclusion is again equivalent to

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} f(x) + \left\langle A^T\lambda^k, x \right\rangle$$

(note that the argument of the minimum exists and is unique due to the strong convexity of f). We finally obtain the following iterations, called the *Lagrange multipliers method*:

$$\begin{aligned}x^{k+1} &= \arg\min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda^k) \\ \lambda^{k+1} &= \lambda^k + \gamma A x^{k+1} \, .\end{aligned}$$

We have the following convergence result.

Theorem 6.1.1. *Assume that f is μ-strongly convex. Assume that the Lagrangian function $\mathcal{L}$ has a saddle point. Set $0 < \gamma < \frac{2\mu}{\|A\|^2}$ where $\|A\|$ is the spectral norm[1] of A. Then, the sequence (x^k, λ^k) generated by the Lagrange multipliers method converges to a saddle point of $\mathcal{L}$.*

Proof. By Lemma 6.1.1, the dual function Φ is differentiable and $\nabla\Phi$ is Lipschitz continuous. It is not difficult to show that the corresponding Lipschitz constant is upper bounded by $\frac{\|A\|^2}{\mu}$. Therefore, the gradient descent on the (negative) dual function yields a sequence λ^k converging to a dual solution. It remains to show that x^k converges to a primal solution $\lambda^\star$. Recall that $x^{k+1} = \nabla f^*(-A^T\lambda^k)$. By continuity of ∇f^*, x^k converges to $x^\star = \nabla f^*(-A^T\lambda^\star)$. Using once again the Fenchel-Young property 4.1.2 and Fermat's rule, it follows that $x^\star = \arg\min_x \mathcal{L}(x, \lambda^\star)$. Thus, $x^\star$ is primal-optimal and $(x^\star, \lambda^\star)$ is a saddle point of $\mathcal{L}$. □

[1] the square root of the largest eigenvalue of A^TA

Remark 6.1.2. The Lagrange multipliers method can be used under slightly milder assumptions. The assumption that f is strongly convex can be replaced by the assumption that the function $y \mapsto \inf\{f(x) \,:\, x \text{ s.t. } Ax = y\}$ is strongly convex. The latter condition is indeed necessary and sufficient to ensure that the dual function Φ is differentiable with a Lipschitz continuous gradient. In the absence of strong convexity assumption on f, note that the quantity x^k may no longer be uniquely defined and the conclusions of the theorem should thus be weakened.

6.1.3 Application: a splitting method

In this paragraph, we instanciate the Lagrange multipliers method as a way to solve the following problem:

$$\min_{x \in \mathcal{X}} f(x) + g(Mx) \tag{6.1.2}$$

where $f : \mathcal{X} \to (-\infty, +\infty]$, $g : \mathcal{Y} \to (-\infty, +\infty]$ are proper closed convex functions and $M : \mathcal{X} \to \mathcal{Y}$ is a linear operator. The above problem writes equivalently as

$$\min\{F(y) \,:\, y \in \mathcal{X} \times \mathcal{Y},\, Ay = 0\}$$

where the function F is defined on $\mathcal{X} \times \mathcal{Y} \to (-\infty, +\infty]$ by $F(x, z) = f(x) + g(z)$ and where $A : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ is the linear operator $A = [M, -I]$ where I denotes the identity, that is

$$A \begin{pmatrix} x \\ z \end{pmatrix} = Mx - z$$

for every $(x, z) \in \mathcal{X} \times \mathcal{Y}$. Provided that the conditions of application of the Lagrange multipliers method are in force, the latter writes:

$$\begin{aligned} (x^{k+1}, z^{k+1}) &= \arg\min_{(x,z)} f(x) + g(z) + \left\langle \lambda^k, Mx - z \right\rangle \\ \lambda^{k+1} &= \lambda^k + \gamma(Mx^{k+1} - z^{k+1})\,. \end{aligned}$$

As a remarkable feature, the first update equation above reduces to solving a minimization problem which is separable in (x, y). Finally, the algorithm reformulates as

$$\begin{aligned} x^{k+1} &= \arg\min_{x} f(x) + \left\langle \lambda^k, Mx \right\rangle \\ z^{k+1} &= \arg\min_{z} g(z) - \left\langle \lambda^k, z \right\rangle \\ \lambda^{k+1} &= \lambda^k + \gamma(Mx^{k+1} - z^{k+1})\,. \end{aligned}$$

The algorithm is called a splitting method, which should be understood in the following sense. The first update equation involves only function f whereas the second one involves only g. The algorithm has an interest in the where f and g are tractable functions and could be separately handled, but the sum $f + g \circ M$ is difficult to minimize.

6.2 Augmented Lagrangian Method

We now apply the proximal point algorithm to the minimization of the (negative) dual function $f^* \circ (-A^T)$. This yields

$$\lambda^{k+1} = \mathrm{prox}_{\gamma f^* \circ (-A^T)}(\lambda^k)\,.$$

In the sequel, we denote by $\mathcal{L}_\gamma$ the *augmented Lagrangian* defined by

$$\mathcal{L}_\gamma(x, \lambda) = f(x) + \langle \lambda, Ax \rangle + \frac{\gamma}{2}\|Ax\|^2\,.$$

Theorem 6.2.1. *The proximal point iteration* $\lambda^{k+1} = \mathrm{prox}_{\gamma f^* \circ (-A^T)}(\lambda^k)$ *writes equivalently*

$$\begin{aligned} &\textit{Find } x^{k+1} \in \arg\min_x \mathcal{L}_\gamma(x, \lambda^k) \\ &\textit{Set } \lambda^{k+1} = \lambda^* + \gamma A x^{k+1}\,. \end{aligned}$$

Proof. By the definition of the proximity operator, the above equation reads equivalently

$$0 \in \gamma \partial(f^* \circ (-A^T))(\lambda^{k+1}) + \lambda_{k+1} - \lambda^k$$

or equivalently

$$0 \in -\gamma A \partial f^*(-A^T \lambda^{k+1}) + \lambda_{k+1} - \lambda^k\,.$$

This means that there exist some $x^{k+1} \in \partial f^*(-A^T\lambda^{k+1})$ such that $\lambda^{k+1} = \lambda^k + \gamma A x_{k+1}$. By the Fenchel Youn property again, the inclusion $x^{k+1} \in \partial f^*(-A^T\lambda^{k+1})$ reduces to $-A^T\lambda^{k+1} \in \partial f(x^{k+1})$ or equivalently

$$0 \in \partial f(x^{k+1}) + A^T(\lambda^k + \gamma A x_{k+1})\,.$$

By Fermat's rule, this is again equivalent to $x^{k+1} \in \arg\min_x \mathcal{L}_\gamma(x, \lambda^k)$. □

As the Augmented Lagrangian Method coincides with a proximal point algorithm on the (negative) dual function, it follows that the sequence λ^k converges to a solution to the dual problem. We remark that the variable x^{k+1} is not necessarily uniquely defined because the arg min may not be unique.

Remark 6.2.1. Apply the Augmented Lagrangian Method to the example (6.1.2). We obtain

$$\begin{aligned} (x^{k+1}, z^{k+1}) &= \arg\min_{(x,z)} f(x) + g(z) + \left\langle \lambda^k, Mx - z \right\rangle + \frac{\gamma}{2}\|Mx - z\|^2 \\ \lambda^{k+1} &= \lambda^k + \gamma(Mx^{k+1} - z^{k+1})\,. \end{aligned}$$

The minimization problem is no longer separable due to the presence of a new quadratic term. In that sense, the Augmented Lagrangian Method cannot be used as a splitting method.

6.3 Alternating Direction Method of Multipliers (ADMM)

The ADMM can be seen as a variant over the augmented method of multipliers, which combines the good features of the later (there is no strong convexity assumption, no restriction on the step size) and the standard method of multipliers (it is a splitting method allowing to "separate" f and g in (6.1.2)). The iterations are given by

$$\begin{aligned} x^{k+1} &\in \arg\min_x f(x) + \left\langle \lambda^k, Mx \right\rangle + \frac{\gamma}{2}\|Mx - z^k\|^2 \\ z^{k+1} &= \arg\min_z g(z) - \left\langle \lambda^k, z \right\rangle + \frac{\gamma}{2}\|Mx^{k+1} - z\|^2 \\ \lambda^{k+1} &= \lambda^k + \gamma(Mx^{k+1} - z^{k+1})\,. \end{aligned} \tag{6.3.1}$$

We remark that in the above iterations, the quantity x^k may, in certain circumstances, not be uniquely defined. However, when M is injective or when f is strongly convex, the arg min in (6.3.1) is unique, and x^{k+1} is unambiguously defined. Therefore, in Equation (6.3.1), the symbol "$\in$" can be replaced by "$=$".

The proof of the following result is taken for granted. We refer to Boyd et al. (2011) for a convergence proof.

Theorem 6.3.1. *Consider a sequence (x^k, z^k, λ^k) satisfying the ADMM iterations and assume that a saddle point of the Lagrangian exists. Then,*

- *the sequence λ^k converges to a solution to the dual problem,*
- *any limit point of the sequence x^k is a primal solution,*
- *the sequence $f(x^k) + g(z^k)$ converges to the primal value $p = \inf f + g \circ M$,*
- *the sequence $Mx^k - z^k$ tends to zero.*

Moreover, if M is injective, then x^k converges to a primal solution.

Bibliography

Bauschke, H. H. and Combettes, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. Springer. 11

Borwein, J. and Lewis, A. (2006). *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer. 28

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122. 35

Boyd, S. and Vandenberghe, L. (2009). *Convex optimization*. Cambridge university press. 6, 29

Brezis, H. (1987). Analyse fonctionnelle, 2e tirage.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2012). *Fundamentals of convex analysis*. Springer Science & Business Media. 32

Nesterov, Y. (2004). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer. 6, 17

Rockafellar, R. T., Wets, R. J.-B., and Wets, M. (1998). *Variational analysis*, volume 317. Springer. 10